PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 19/00 — A1

(11) International Publication Number: WO 97/27560

(43) International Publication Date: 31 July 1997 (31.07.97)

(21) International Application Number: PCT/US97/01727

(22) International Filing Date: 29 January 1997 (29.01.97)

(30) Priority Data:
08/593,446 29 January 1996 (29.01.96) US

(71) Applicant: ONCORMED [US/US]; 205 Perry Parkway, Gaithersburg, MD 20877 (US).

(72) Inventor: EVANS, Steven; 418 North 38th Street, Omaha, NE 68131 (US).

(74) Agents: JONDLE, Robert, J. et al.; Rothwell, Figg, Ernst & Kurz, 555 13th Street, N.W. # 701 East, Washington, DC 20004 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHODS FOR IDENTIFYING HUMAN HEREDITARY DISEASE PATTERNS

(57) Abstract

The present invention relates to a method of determining human hereditary disease risk factors, a method of determining hereditary cancer patterns presenting in cancer family histories, and a method of determining whether a cancer family history represents a hereditary pattern.

TITLE OF THE INVENTION

"METHODS FOR IDENTIFYING HUMAN HEREDITARY DISEASE PATTERNS"

BACKGROUND OF THE INVENTION

The present invention relates to a method of determining human hereditary disease risk factors, a method of determining hereditary cancer patterns presenting in cancer family histories, and a method of determining whether a cancer family history represents a hereditary pattern.

Recognition of the role of heredity in cancer has increased steadily over time and is discussed in Lynch et al. *Genetic Epidemiology of Cancer*, Boca Raton: CRC Press, (1989). For example, Lynch et al., *Surg. Clin. Of NA*, 70:753-774, (1990) has estimated that approximately 9% of carcinoma of the breast is consistent with hereditary breast cancer, and an additional 15-20% will be clearly familial. In addition, the potential for gene testing to confirm hereditary cancer has escalated in recent years, with the discovery of genes that cause breast cancer, colon cancer, and several other cancer syndromes. However mass testing of the population with dozens of expensive gene tests is not viable. The most effective way to identify hereditary cancer families has been with a detailed and accurate cancer family tree. It has been noted, that "[a] thoroughly compiled family history of cancer harbors the potential of being the most cost beneficial component of the patient's workup" (Lynch, HT, "Cancer and the family history trail, *New York State Journal of Medicine*, pp. 145-147, April, 1991). From

SUMMARY OF THE INVENTION

The present invention relates to a method of determining human hereditary disease risk factors, a method of determining hereditary cancer patterns presenting in cancer family histories, and a method of determining whether a cancer family history represents a hereditary pattern implemented on a computer.

More specifically, the invention relates to a method for determining the existence of a hereditary disease risk in a patient, comprising the steps of: compiling in a computer a database made up of a plurality of records each pertaining to an individual and containing a history of at least one specific disease in a family of that individual, with a plurality of parameters relating to each family member identified in the history; defining a plurality of functions each pertaining to one of the parameters and assigning predetermined weights to the functions based on values of the parameters; for each record in the database, summing the weights obtained for each of the functions to obtain a total value for each of the functions, identifying the record as presenting a hereditary pattern if the total value is above a predetermined threshold, and grouping the identified record into a subset of records; for each record in the database, applying expert knowledge generated rules to independently identify records as presenting hereditary patterns; comparing the independently identified records with the subset of records, and validating defined functions if a predetermined minimum percentage of records in the subset are consistent with the independently identified records; and using validated functions as a recognizer of hereditary disease patterns in a family history of the patient.

The invention also relates to a method wherein the plurality of records contain histories of instances of cancer including breast, ovarian, endometrial, prostate, maliginant melanoma and colon cancer.

The present invention further relates to a method including the step of assigning weights to particular attributes used in the functions, and defining attributes as significant in the definition of specific hereditary disease patterns if a minimum percentage of records in the subset are consistent with the independently identified records.

which, if any, are the "example cases" of hereditary cancer in the data in order that one might obtain a logical rule set from data mining methods.

Expert Rule: The term expert rule is intended to mean a set of rules, each of which is comprised of logical measurable conditions, in mathematics, called a testable predicate which is an expression or condition that can be determined to be true or false based on measurement or data, wherein these conditions are joined by logical operators, including the operators AND, OR, and NOT, having the usual logical meaning. Given an example of the pattern about which the rule is concerned, one can examine the measurable characteristics of a pattern and ascertain whether it satisfies all the conditions of one or more of the rules in an expert rule set.

Recognizer: The term recognizer is intended to mean a rule set for a class of patterns for which any member of the class satisfies one or more of the rules in the rule set. For example, a set of reasonable conditions might be:

a) for a specific person "p", there was a colon cancer in a first-degree relative of that person,

b) for a specific person "p", there was a colon cancer occurring before age 35 in a first-degree family member relative

c) for a specific person "p", there were three or more cases of colon cancer in first or second-degree relatives

And one rule might be: (a) AND ((b) OR (c))

The meaning of this rule is that if a person had a first-degree relative with colon cancer and that person had either a first degree relative with colon cancer occurring before the age of 35 or had three or more cases of colon cancer in first or second degree relatives, then that person's pattern meets or satisfies the rule. If this is one of the rules for a recognizer of hereditary colon cancer, then the person's pattern is said to be classified as a member of the recognized pattern set i.e., the pattern is a hereditary colon cancer pattern. If any member of the class of patterns constituting hereditary colon cancer were to fit the rule above, the rule above would then be the rule set or recognizer of hereditary cancer, although obviously the set usually has many different rules in

the nature of hereditary cancer patterns. Core genetics principles are employed which tend to characterize cases that would be good candidates for representing hereditary cancer. For example, one such principle is that typically hereditary cancer arises much earlier in the patient's life than does non-hereditary cancers. The application of a set of key principles will divide the set of all cases into those which reflect some mix of these principles, beyond a threshold, and those that do not. The degree to which the principles apply in a case is the extent to which a case tends to be a candidate to represent a hereditary pattern. All such cases that exceed some measurable threshold will be declared confirmed candidates for representing a possible hereditary pattern.

A second stage of the present invention is to take the cases that are confirmed candidates and then determine both the genetic principles which classify the confirmed candidates as well as the patterns of cancers in these cases which characterize the candidates. Specifically two sets of rules are determined which classify the cases into various composite pattern sets wherein each rule describes one or more patterns. This stage uses software called data mining techniques in order to obtain computer-generated rules which group or classify the candidate cases, using first the genetic principles and then the cancer family history patterns. In the first instance, the genetic principles characterize the candidate cases by a rule set in which the use of a genetic principle in the set indicates that this principle is a factor in the overall determination of hereditary cancer for the index cancer designated. These second rule set is a refining definition based on cancer patterns for the candidate cases. Each rule in the second set is examined, and the specific cancers used in each rule become the cancers which characterize the index cancer, e.g., endometrial cancer is a coincident cancer in rules that characterize colon cancer.

In the third stage, these coincident cancers are combined with the candidate-defining genetic principles to create the complete recognizer which is a composite set of clinically oriented defining rules.

One method of the present invention includes the following steps:

Step 1. The first step is to obtain a database which contains the cancer family

Step 3. Once the individual cancer family histories are expanded into many different cancer family patterns, a specific cancer or cancers, of interest to be recognized is specified, e.g., breast cancer, and all entries in the newly expanded database of patterns which are indexed by the selected cancer are retrieved. Although several cancers may be picked at once with multiple cancers being an additive case, we shall give this description as if only one index case was selected, merely for ease of explanation.

Step 4. Next, a further subset of all the cancer histories concerned with the cancer of interest is selected which shall represent candidate hereditary cancer cases from which common patterns are extrapolated. This is accomplished by positing a set of descriptors or genetic principles which typically are known to be involved in hereditary disease. One such descriptor is "early-age-of-onset," since hereditary cancers tend to occur earlier than normal. Each descriptor is quantified. For example, if a cancer family history has a cancer occurring before the age of 50, one point is assigned; if there is a cancer occurring below the age of 45, two points are assigned; and if one occurs before the age of 35, three points are assigned. In the instant invention five such descriptors have been employed although the method in general could be instantiated with as many principles as desired. These genetic principles can include (1) early age of onset, (2) presence of cancer over several generations, (3) several cancers in the same generation, (4) multiple cases of the same cancer occurring, and (5) a high proportion of the relatives in the family expressing cancer. Each is quantified assigning increasing points the more the principle is expressed in the case history. Cases with a certain total score or higher are then labeled as candidates for the group that represents a hereditary cancer pattern and are then selected for further analysis. Although the principles have been quantified, the actual point

expressions in the family as defining characteristics. An arithmetic calculation is then made to determine what percent of all the "true" cases (i.e., the percent of the candidate set from step 4) meet any one rule. A rule that is met by at least 5% of all true cases, which is then defined as a significant rule, is examined, and all attributes in such significant rules are listed and are determined to be significant risk factors for the index cancer.

The attributes that are significant from the genetics principles list are the genetic principles which are pertinent to the selected index cancer and its pattern of presentation, while the cancers which are listed in the cancer-patterns significant rules are the pertinent cancers for the selected index cancer in question.

The outcome or output from step 5 is two sets of attributes which are comprised of entries from the genetic principles list and the cancers list. The aggregate of the attributes found in the rules that meet the 5% rule are defined to be the correlated genetic principles and cancer patterns which are coincident with and can characterize the selected index cancer pattern under consideration.

Step 6. Finally, data mining software is run again, marking as of interest the enumerated significant genetic principles and the significant cancer patterns from step 5 which shall characterize the "true" set in order to get a combined rule set of both genetic principles and cancer patterns, which characterizes the "true" set. The software derives a set of rules which uses the composite attributes that have been derived by this process (as per step 5), and which characterize the "true" set as defined. The set of all rules in this step 6 which are significant (i.e., 5% or more of the "true" set meet the rule) constitute the hereditary cancer rules which recognize the hereditary cancer pattern for the index cancer selected. Thus the set of all such rules in this stage is the hereditary cancer recognizer for the

Examples 5-11.

6. Experiments and refinements of the process as well as evaluating the recognizers for usefulness have been conducted as shown in Examples 12-17.

For example, focusing on colon cancer, one process of the present invention yields a set of rules (a recognizer) so that 100 % of all the cases in the independent files available to the experimenters, labeled hereditary to date by medical experts are correctly labeled by the recognizer.

We determined the nature and extent of the hereditary colon cancer cases which are not recognized, if any, in further experimentation and how to adjust the process so that the automatically constructed recognizer has greater selectivity and sensitivity. We have also replicated this effort to analogous recognizers for breast cancer, malignant melanoma, and pancreatic cancer.

EXAMPLES

The following examples are provided to further illustrate the present invention and are not intended to limit the invention beyond the limitations set forth in the appended claims.

Example 1

Developing a Database

A database of clinical cases was converted into an information source that permits the development of appropriate rules. The database was obtained from the Hereditary Cancer Institute (HCI) of the Creighton University School of Medicine, Omaha, Nebraska. This database had cases describing patient cancer family data. There was not an independent assessment in the database which validated that a patient in the database is in fact a hereditary-cancer-affected person. The database represented people who contacted HCI, and so we believed there were some hereditary families in the database. However this database did not have a hereditary cancer assignment of "true" or "false" regarding each case (in which true = a hereditary cancer carrier), although such an assignment is needed for either traditional neural network or data mining methodology.

Example 3

Defining Genetic Principles

We presented the five basic broad-level genetic principles that may apply in any hereditary pattern, which were:

a. Inheritance may show up more than once in the same generation (horizontal or generational inheritance). This principle will be made precise as a function f_1 in the application example below and is abbreviated GENLINE in the computer output.

b. Inheritance may show up from the prior generation to a subsequent generation (which we call inter-generational or vertical inheritance). This principle will be made precise as a function f_2 in the application example below and is abbreviated VERLINE in the computer output.

c. There may be numerous instances of various (different) manifestations of the pattern over multiple generation (general intensity). This principle will be made precise as function f_3 in the application example below and is abbreviated INTENSITY in the computer output.

d. There may be numerous instances of the same manifestation of the pattern over multiple generations so that there is a specific intensity of some specific manifestation. This principle will be made precise as function f_4 in the application example below and is abbreviated SPECINTEN in the computer output.

e. Since it is genetically based, it occurs sooner (early age of onset) in the age of the organism expressing the pattern (i.e., as soon as the genes can begin expressing themselves) rather than at later ages such as patterns caused by external (i.e., non-genetic events whose probability of occurrence can build over time and thus more and more likely occur as more and more time passes (late age of onset). This principle will be made precise as a function f_5 in the application example below and is abbreviated EARLY in the computer printout.

$f_4(\bar{x}^k) = 3$ if there are more than 3 identical cancers on the same side.

$f_5(\bar{x}^k)$: Early age of cancer onset - this attribute states that early age of onset of cancer is an interesting attribute. We assigned:

$f_5(\bar{x}^k) = 1$ for each cancer diagnosed by the age of 50.

$f_5(\bar{x}^k) = 2$ for each cancer diagnosed by the age of 45

$f_5(\bar{x}^k) = 3$ for each cancer diagnosed by the age of 35.

Using the five principles or functions above, we apply each function to each case record to assign points depending on the data contained in each record of the case under review. The more points, the more the case is very serious in that it is reflecting an example of a hereditary inheritance. We add the five functional values together to get a total sum (and we may multiply any one of the functions times a weight if we wish to emphasize that attribute more than another. A specific point total is selected, and for all cases of the index cancer that is selected (e.g., colon cancer proband cases) that accrue these many points or more, we label this the defining subset from the larger set which will be used to define the recognizer. This set of cases whose point total is above a threshold is the "true" set for purposes of using standard data mining algorithms to characterize the cases in the set. Formalistically, for our colon cancer recognizer, we assign the value "true" to a case x in the database if:

$$\sum_{i=1}^{5} a_i f_i(x) \geq v$$

for a threshold value v and for a_i where a_i is any numerical weight for f_i.

Example 4

As noted in Example 2, a database is obtained and the individual cases permuted so that variations in patterns are obtained. As per Example 3, each individual patient history (or case) is evaluated in terms of how it meets each of the hereditary principles (in the instant case, by a computer program). Example 3 shows how, for an instant application, numeric values are provided with an assignment strategy so that a case accrues points depending on how many cancers there are in the family, the type and extent, etc. All cases that pass a selected threshold value in Example 3 are the candidate

Applying the 5% significance rule, rule 11 is omitted, and the rest remain. Inspecting the attributes (the hereditary principles) which occur in the 10 viable rules, we find all but "INTENSITY" are used. INTENSITY occurs only with a <>sign in all rules and since one of the viable choices is "strictly less than 1", the value of zero is a choice, meaning zero INTENSITY is a permissible condition (i.e., no INTENSITY). So this attribute falls away, and the remaining are significant.

This procedure is repeated using all of the cancers that occur in the family histories of all the cases. The results are again presented below from the same software package. In this run, only the twelve rules 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 13 and 14 had more than 51 cases meeting the specific rule.

```
1  |    |  |[CA14_A<1] & [4<= CA1_A <=14] & [CA1_B>=2]
   | OR
2  |    |  |[CA1_A>=6] & [1<= CA1_B <=7]
   | OR
3  |    |  |[4<= CA1_A <=14] & [1<= CA1_B <=7] & [CA2_A>=1]
   | OR
4  |    |  |[4<= CA1_A <=14] & [CA9_B>=1]
   | OR
5  |    |  |[CA2_A>=1] & [CA9_B<=0 or CA9_B>=2] & [CA2_B>=1]
   | OR
6  |    |  |[CA14_A<1] & [4<= CA1_B <=7]
   | OR
7  |    |  |[2<= CA1_B <=7] & [CA2_A<=0 or CA2_A>=2] & [CA2_B>=1]
   | OR
8  |    |  |[4<= CA1_A <=14] & [CA2_B>=1]
   | OR
9  |    |  |[4<= CA1_A <=14] & [CA2_A>=2] & [CA24_A<=0 or CA24_A>=3]
   | OR
10 |    |  |[CA2_A>=3]
   | OR
11 |    |  |[CA1_A>=8] & [CA2_A>=1]
   | OR
12 |    |  |[CA42_A>=1] & [1<= CA1_B <=7]
   | OR
13 |    |  |[CA1_A<=3 or CA1_A>=8] & [CA1_B>=3] & [CA2_A<=0 or CA2_A>=2] & [CA9_B<=0 or
   |    |  [CA9_B>=2]
   | OR
14 |    |  |[CA1_A>=6] & [CA24_A>=1]
   | OR
15 |    |  |[CA1_B=2] & [CA2_A>=1] & [CA9_B<=0 or CA9_B>=2]
   | OR
16 |    |  |[CA3_A<1] & [1<= CA1_B <=7] & [CA2_A>=1] & [CA6_B>=1]
   | OR
17 |    |  |[CA2_A>=2] & [CA6_B>=1]
   | OR
18 |    |  |[CA6_B>=1] & [CA24_A>=1]
   | OR
19 |    |  |[CA3_A>=1] & [4<= CA1_A <=14] & [CA2_A<=0 or CA2_A>=2]
   | OR
20 |    |  |[CA3_A>=1] & [CA1_A>=6]
   | OR
21 |    |  |[CA1_A<=3 or CA1_A>=8] & [CA2_A>=2] & [CA24_A>=1]
   | OR
22 |    |  |[CA1_B=2] & [CA9_B>=1] & [CA2_B<=0 or CA2_B>=3]
   | OR
23 |    |  |[CA3_A<1] & [CA18_A>=1] & [CA6_B>=1]
   | OR
24 |    |  |[4<= CA1_A <=14] & [CA2_A>=1] & [CA24_A>=1]
```

permissible rules. Data mining algorithms let you specify what attributes you wish to use in classifying the "true" cases, and since we have derived in the two steps above the key elements, we are ready for the final run. The results are given below:

```
1  |    |  |[EARLY>=6] & [SPECINTEN>=3]
   | OR
2  |    |  |[GENLINE=3] & [EARLY>=6] & [SPECINTEN>=2]
   | OR
3  |    |  |[GENLINE=3] & [EARLY>=6] & [VERLINE>=1]
   | OR
4  |    |  |[EARLY>=9]
   | OR
5  |    |  |[EARLY>=6] & [SPECINTEN>=2] & [VERLINE>=1] & [CA1_B<=0 or CA1_B>3]
   | OR
6  |    |  |[EARLY>=6] & [VERLINE>=2]
   | OR
7  |    |  |[GENLINE=3] & [EARLY>=3] & [SPECINTEN>=3] & [VERLINE>=2]
   | OR
8  |    |  |[EARLY>=3] & [CA1_A>=6] & [VERLINE>=1]
   | OR
9  |    |  |[GENLINE=3] & [EARLY>=3] & [4<= CA1_A <=9] & [VERLINE>=1] & [1<=CA1_B <=5]
   | OR
10 |    |  |[GENLINE=3] & [EARLY>=3] & [SPECINTEN>=3] & [CA1_A<=3 or CA1_A>=8] &
   |    |  |[CA2_A>=1]
   | OR
11 |    |  |[GENLINE=3] & [EARLY>=3] & [VERLINE>=2] & [1<= CA1_B <=5] & [CA2_A<=0 or
   |    |  |CA2_A>=2]
   | OR
12 |    |  |[EARLY>=3] & [SPECINTEN>=3] & [CA2_A>=2]
   | OR
13 |    |  |[EARLY>=3] & [CA1_B>=2] & [CA2_A>=1]
   | OR
14 |    |  |[GENLINE=3] & [EARLY>=3] & [CA1_B=2]
   | OR
15 |    |  |[EARLY>=3] & [8<= CA1_A <=9]
   | OR
16 |    |  |[GENLINE>3] & [4<= CA1_A <=9] & [VERLINE>=2] & [CA2_A>=1]
   | OR
17 |    |  |[SPECINTEN>=3] & [CA1_A<=3] & [VERLINE>=2] & [CA1_B<=0 or CA1_B>=3] &
   |    |  |[CA2_A<=0 or CA2_A>=2]
   | OR
18 |    |  |[EARLY>=3] & [CA1_B>=6]
   | OR
19 |    |  |[GENLINE<>3] & [SPECINTEN>=3] & [VERLINE<=1] & [CA2_A>=1]
   | OR
20 |    |  |[GENLINE<>3] & [EARLY>=3] & [SPECINTEN>=2] & [VERLINE>=2] & [CA1_ B<=0 or
21 |    |  |[EARLY>=3] & [SPECINTEN<=1] & [VERLINE>=2] & [CA1_B<=0 or CA1_B>= 3] &
   |    |  |[CA2_A>=1]
```

We again apply the 5% rule to get the final, significant rule set. All rules taken together characterize more than 93% of all of the true cases, indicating that this rule set very strongly classifies the true set. Looking at the cases that meet each rule, the first 16 rules meet the 5% cutoff, and the remaining do not. Inspection of the cancers which occur in these 16 rules yields just two: CA1 and CA2 (colon and endometrial cancer). Since this is the colon cancer recognizer, it is a tautology that added colon cancers will define a hereditary condition. Thus the next, non-colon cancer occurring, endometrial

of "early age presentations" requires either one breast cancer before the age of 35 with another before the age of 45; another acceptable "early age presentation" would be two before the age of 45 and one before the age of 50. It is also important to observe that each rule derived is in effect a family of rules with a wide number of variations which are precisely specified for different values of variables in each rule.

Example 6

Predicted Syndromes

Associated syndromes for each cancer type t are summarized below. For each cancer type t, the associated cancers that are significant in support of a hereditary pattern are listed in Table 2.

Table 2

Colon	**Melanoma**	**Pancreatic**
endometrial/uterine	small intestine	ovarian
kidney	cervix	stomach
stomach	lung	brain
pancreas	stomach	lung
lung	lip	tongue
ovarian	urinary bladder	prostate
	pancreatic	lip
Breast-maternal	**Breast-paternal**	**Ovarian**
ovarian	ovarian	breast
colon	prostate	endometrial/uterine
prostate	endometrial/uterine	colon
endometrial/uterine	colon	lung
	lung	prostate
	stomach	

The most highly correlated cancers for various cancer types are summarized below in Table 3.

Table 3

Cancer Type	Key Correlated Cancer
colon	endometrial
breast-maternal	ovarian
breast-paternal	ovarian

colon cancer (all on the maternal side)

Or

One 1st-degree breast cancers & one 2nd-degree prostate cancer (all on the maternal side)

Example 9

Breast - paternal side

A person with breast cancer has a hereditary cancer pattern if there are additionally at least the following:

Three 1st-degree breast cancers & one 2nd-degree breast cancer (all on the paternal side)

Or

Two 1st-degree breast cancers & one 2nd-degree ovarian cancer (all on the maternal side)

Example 10

Pancreatic Cancer

A person with pancreatic cancer has a hereditary cancer pattern if there are additionally at least the following:

Two 1st-degree pancreatic cancers & two 2nd-degree pancreatic cancers

or

Three 1st-degree pancreatic cancers & one (either 1st-degree tongue or 1st-degree melanoma)

Or

Two 1st-degree brain cancers

or

One 2nd-degree stomach cancer

Example 11

Melanoma

A person with melanoma has a hereditary cancer pattern if there are additionally at least the following:

outcome.

The process of automated recognizer rule set construction was assessed and validated with *a priori* knowledge which was available for hereditary colon cancer. Similarly the validation for breast cancer, melanoma, and ovarian cancer was repeated. Through this assessment process, the automated recognizer development process was refined and was shown to produce a powerful recognizer development technique. Further validation was achieved by running the recognizer against prior cancer cases for which we had Dr. H. T. Lynch's diagnosis regarding their hereditary nature.

This entire design process was also generalized by positing a universe of genetic principles for all the cases of cancer, and then letting the rule hypothesis component posit rules (i.e., a selection from the genetic principles) which characterize cases of hereditary cancer. For example, for every cancer case, the race, hair color, and so forth along with the age of onset, the number of cancers. and other possible genetic principles were itemized. Then using the exact same process, the rules classify the hereditary cancer cases by pertinent genetic principles such as "cancer patients tend to be younger," "have more instances in their family," and so forth. Thus the pertinent genetic principles from among all principles which could be speculated are identified using the same process as presented above.

Example 13

Study Evaluating Significant Factor Identification

This study compared the responses of Dr. Henry T. Lynch and Dr. Stephen J. Lemon of HCI to a set of factors isolated by the Recognizer in its analysis of hereditary colorectal cancer (HCRC) cases. This set of outcomes corresponds to key factors the Recognizer identifies as possibly relevant to incorporate into its decision-making logic to recognize HCRC. These factors are defined by the Recognizer as useful in its efforts to construct a set of rules to define HCRC.

A list of 42 cancers was provided to Dr. Lynch and Dr. Lemon, and each was asked separately to mark a ranking using the scale for each cancer as shown in Table 5. The list of cancers was taken from the list which encoded the database used to create the

liver/intrahepatic, gallbladder/bile duct/ampulla of Vater, and appendix). The Recognizer and Dr. Lemon were also in close agreement on two other choices ("definitely" versus "probably" on ovarian and cervix). Hepatobiliary cancer did not specifically appear in the database as a coded choice, although Dr. Lynch and Dr. Lemon both had wished to include it as another choice. The Recognizer did not select either Hepatobiliary-related entry which was available to it (liver/intrahepatic or gallbladder/bile duct/ampulla of Vater) although each was selected by either Dr. Lynch or Dr. Lemon."

The Recognizer was in accord in six of seven cases (86%) on HCRC-related cancers for which Drs. Lynch and Lemon were in accord (ranking of 1 or 2). The system essentially considered 35 cancers as irrelevant, and of these, Drs. Lynch and Lemon were in mutual concurrence, ranking each 3 or 4, for 29 of them (81%). If one considers the frequency of occurrence of the associated cancers for the cancers correctly identified by the Recognizer, (i.e., those for which there was agreement between Dr. Lynch and Dr. Lemon), the Recognizer defined critical attributes for its focus most of the time.

There is a second use of such results by the Recognizer over and above automated HCRC detection. There is clinically diagnostic value in the discovery of associated cancers which help confirm a hereditary pattern. In addition, the identification of associated cancers helps guide future gene discovery efforts, as well as assists in the proper interpretation of gene-testing results.

This first stage evaluation of initial factors selected by the Recognizer suggested that use of such Recognizers is applicable over a wide sampling of typical cases with a very high level of sensitivity in its recognition capability.

Example 14

An Evaluation of Patterns

The purpose of this study was to ascertain how well the Recognizer created valid patterns of HCRC which it uses to evaluate a case.

Drs. Lynch and Lemon were asked to indicate agreement or disagreement with a list of 16 clinical patterns to the extent that each pattern would permit the designation of HCRC for an individual presenting with such a pattern. Each gave an independent

Table 7

Would the Following Pattern permit the Conclusi n f HCRC?	Dr. Lynch's Resp nse	Dr. Lemon's Response
#1	Yes	Yes
#2	Yes	Yes
#3	Yes	Yes
#4	Yes	Yes
#5	Yes	Yes
#6	Yes	Yes
#7	Yes	Yes
#8	Yes	Yes
#9	Yes	Yes
#10	Yes	Yes
#11	Yes	Yes
#12	Yes	Yes
#13	Yes	Yes
#14	Yes	Yes
#15	Yes	Yes
#16	Yes	Yes

Drs. Lynch and Lemon agreed with all 16 patterns the Recognizer had created. The Recognizer's final output was highly sensitive in detecting HCRC.

Table 8

1. All on the same side of the family:
   a. the proband has colon cancer, and
   b. there are more than 3 identical cancers (either colon or endometrial), and
   c. cancer(s) (colon or endometrial) with early onset total a minimum of 6 points
   (where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1 point)
2. All on the same side of the family:
   a. the proband has colon cancer, and
   b. there are 3 or more of the same cancer (either colon or endometrial) in the same generation, and
   c. there are 2 or more identical cancers, and
   d. cancer(s) (colon or endometrial) with early onset total a minimum of 6 points
   (where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1 point)

c. there are 4 or more first degree colon cancers, and

d. there is 1 or more second degree colon cancer, and

e. cancer(s) (colon or endometrial) with early onset total a minimum of 3 points

(where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1

10. All on the same side of the family:

a. the proband has colon cancer, and

b. there are 3 or more of the same cancer (colon or endometrial) in the same generation, and

c. there are 3 or more first degree colon cancers, and

d. there is 1 or more first degree endometrial cancer, and

e. cancer(s) (colon or endometrial) with early onset total a minimum of 3 points

(where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1 point)

11. All on the same side of the family:

a. the proband has colon cancer, and

b. there are 3 or more of the same cancer (colon or endometrial) in the same generation, and

c. there are identical cancers (colon or endometrial) in 2 or more generations, and

d. there is 1 or more second degree colon cancer, and

e. cancer(s) (colon or endometrial) with early onset total a minimum of 3 points

(where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1 point)

12. All on the same side of the family:

a. the proband has colon cancer, and

b. there are more than 3 identical cancers (colon or endometrial), and

c. there are 2 or more first degree endometrial cancers, and

d. cancer(s) (colon or endometrial) with early onset total a minimum of 3 points

(where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1 point)

13. All on the same side of the family:

a. the proband has colon cancer, and

b. there are 2 or more second degree colon cancers, and

c. there is 1 or more first degree endometrial cancer, and

d. cancer (colon or endometrial) with early onset total a minimum of 3 points

(where cancer by age 35 = 3 points; between 36-45 = 2 points; between 46-49 = 1 point)

14. All on the same side of the family:

a. the proband has colon cancer, and

b. there are 3 or more of the same cancer (colon or endometrial) in the same generation, and

c. there are 2 second degree colon cancers, and

Table 9

Colon Cancer Cases				
Case Number	Determination by Dr. Lynch	Confirmation by Recognizer	Recognizer Pattern Match	Additional Findings to Match One Pattern
1	Definite Hereditary Pattern	Yes	1,2,3,4,5,6,7	N/A
2	Definite Hereditary Pattern	Yes	5,6	N/A
3	Definite Hereditary Pattern	Yes	1,4,5,6	N/A

Table 10

Colon Cancer Cases				
Case Number	Determination by Dr. Lynch	Confirmation by Recognizer	Recognizer Pattern Match	Additional Findings to Match One Pattern
1	Putative Hereditary Pattern	No	None	Additional early onset less than age 35 to meet #5
2	Putative Hereditary Pattern	No	None	change of ovarian cancer diagnosis to endometrial with age of onset 2 years earlier to meet #13
3	Putative Hereditary Pattern	No	None	One additional colon cancer in same generation to meet #7
4	Putative Hereditary Pattern	No	None	Age of onset of one of the colon cancers prior to age 35 to meet #7 and #11

All the definite patterns of HCRC (N=3) matched one or more of the Recognizer patterns. None of the four putative HCRC cases matched a Recognizer pattern. However the missing data elements required to match at least one pattern were minimal.

Therefore, the Recognizer gives the correct response for all definite HCRC cases, and gives a negative response in all 4 putative HCRC cases. Thus the Recognizer developed was both sensitive and specific for definite HCRC patterns.

"having thorns" would also; having any kind of fruit growing would be a negative indicator. By a weighted summing of the positive attributes presented and subtracting the negative, a value is obtained, which if greater than a fixed amount, indicates we have a rose. This methodology requires having possible rose attributes determined and the truth value of the assignment for "rose" (or not) given to the system so that it can balance all its weights to signal a rose. In our situation the detailed attribute universe is not known, in general, and more importantly there is no "labeling" of what constitutes a valid rose. Thus neural net methodology is not applicable to the tasks undertaken.

Another limitation is that the result of the application of neural net methodology is a strictly performance-based network, in the form of essentially a mathematical model that predicts set membership. Such networks do not lend themselves to intellectual clinical model validation although the network's performance can be measured by testing. Physicians want to know why a claim is made, not just a numerical calculation that it should be made.

definition of specific hereditary disease patterns if a minimum percentage of records in said subset are consistent with said independently identified records.

5. A method according to claim 1, wherein an additional record is created in said database for each relative of said individual who is identified as having had said disease.

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MEDICAL INFORMATICS, ETHICS, CARDIOLOGY, INSTRUMENTATION, SAN DIEGO, OCT. 28 - 31, 1993, vol. 15 PART 2, 28 October 1993, SZETO A;RANGARAJ M RANGAYYAN, page 646/647 XP000436878 CONSTANTINOU I ET AL: "A DATABASE SYSTEM FOR THALASSAEMIA MUTATIONS" see the whole document ---	1-5